The Modern RNP World of Eukaryotes

LESLEY J. COLLINS, CHARLES G. KURLAND, PATRICK BIGGS, AND DAVID PENNY

Abstract

Eukaryote gene expression is mediated by a cascade of RNA functions that regulate, process, store, transport, and translate RNA transcripts. The RNA network that promotes this cascade depends on a large cohort of proteins that partner RNAs; thus, the modern RNA world of eukaryotes is really a ribonucleoprotein (RNP) world. Features of this “RNP infrastructure” can be related to the high cytosolic density of macromolecules and the large size of eukaryote cells. Because of the densely packed cytosol or nucleoplasm (with its severe restriction on diffusion of macromolecules), partitioning of the eukaryote cell into functionally specialized compartments is essential for efficiency. This necessitates the association of RNA and protein into large RNP complexes including ribosomes and spliceosomes. This is well illustrated by the ubiquitous spliceosome for which most components are conserved throughout eukaryotes and which interacts with other RNP-based machineries. The complexes involved in gene processing in modern eukaryotes have broad phylogenetic distributions suggesting that the common ancestor of extant eukaryotes had a fully evolved RNP network. Thus, the eukaryote genome may be uniquely informative about the transition from an earlier RNA genome world to the modern DNA genome world.

Key words: molecular evolution, origin of eukaryotes, RNA infrastructure, RNA world

Some functions of RNA have been well known for many years, including the roles of mRNA, the ribosome, small nuclear RNAs (snRNAs) in mRNA splicing, small nucleolar RNAs (snoRNAs) in rRNA modification, RNase P in tRNA cleavage, and tRNA in translation. Other discoveries of the past decade have expanded the roles of RNA in the modern eukaryote cell, and it is remarkable that almost all these RNAs are only found in association with proteins. Thus, the modern RNA infrastructure (Collins and Penny 2009) is in eukaryotes an extended ribonucleoprotein (RNP) world.

The rates of catalysis by protein-free RNA are low compared with the rates of fully constituted RNP complexes or compared with proteins themselves (Jeffares et al. 1998). Likewise, in the modern world, the rates of peptide bond formation by nearly protein-free tRNA are very low compared with the rates of fully constituted ribosomes (Polacek and Mankin 2005). Discoveries, such as the core of the ribosome being a ribozyme (Nissen et al. 2000), apply to all 3 domains of life; others such as intron-encoded functional RNAs are now only found in eukaryotes. If we concentrate on eukaryotes, because they have a much richer set of RNA–protein interactions than prokaryotes (see Collins and Penny 2009), then we can consider possible regulatory reactions that may have occurred in an RNP world before DNA. Protein–RNA complexes are interesting in that, although RNA may be directly involved in catalysis, proteins now play an important auxiliary role (Bokov and Steinberg 2009). Accordingly, we focus on the infrastructure of RNP complexes that mediates the gene expression cascade of modern eukaryotes (other reviews on the richness of RNP processes in eukaryotes are available [Bompfunewerer et al. 2005; Mattick 2005; Mattick and Makunin 2006; Rodriguez-Trelles et al. 2006]).

It is not just our knowledge of RNA-based regulation and modification of RNA that has expanded (Collins and Penny 2009). Highly relevant is the acceptance that the eukaryotic cell has a complex structure with high concentrations of macromolecules, approaching the concentration of proteins in crystals (Ellis 2001). The concentration of macromolecules is so high (molecular crowding) that diffusion of macromolecules is strongly limited (Richet et al. 2008); efficient kinetics requires that macromolecules function together in complexes. Strong compartmentalization with macromolecular complexes is essential for any cell, especially for a large cell, to function efficiently (Kurland et al. 2006; Robinson et al. 2007).

Before we can evaluate hypotheses for the origin of eukaryotes, we must identify characteristics of their last common ancestor. We agree with the view that theories for the origin of eukaryotes that are not based on the continuity of evolution are inadequate; but where is that continuity to be found? Alternatively, what features of the last common ancestor can we infer from modern eukaryotes? For example, was RNase MRP (which processes the rRNA...
transcript, see Woodhams et al. 2007) present in the ancestral eukaryote? Similarly, did that ancestor have a full nuclear pore (Alber et al. 2007) such as that seen with the spliceosome complex (Collins and Penny 2005)? And because we now understand that RNA-based processes are central to most RNP complexes, we can also ask how these RNA and protein complexes evolved, directly from the putative RNA world or from a world with associated peptides forming RNPs, or were RNP complexes lost in prokaryotes (and regained in eukaryotes)?

The RNP infrastructure (as an extension of the RNA infrastructure) of modern eukaryote cells may provide a window through which to study the transition from an ancient RNA plus protein world to the modern DNA genome world.

Roles of RNPs in Eukaryotes

The roles of RNA-mediated processes in eukaryote cells are extensive and complex, and form a network (Collins and Penny 2009) where RNA molecules are involved in processing other RNA transcripts (Woodhams et al. 2007). These include the ribozymes RNase P (tRNA cleavage), RNase MRP (rRNA cleavage), and the snRNAs (U1, U2, U4, U5, and U6—cleavage and ligation of mRNA during splicing). The processing of tRNA, rRNA, and mRNA are often investigated individually, but there are cross linkages between them. For example, the snRNAs are involved in splicing mRNA, which can release snoRNAs that identify the modification sites for tRNA during maturation of the ribosome. The involvement of RNA processing other RNA molecules in the RNA infrastructure (Collins and Penny 2009) is certainly thought provoking. Are there reasons that RNA is processing RNA in particular; is this possibly a relic of an earlier RNA genomic world from before DNA took over the coding of genetic information?

We focus first on that part of the complex network of RNP-based interactions that mediate mRNA splicing. The spliceosome associated with major (U2) splicing is intricate (Figure 1), with 5 U snRNAs and (in humans) around 200 proteins. RNA molecules are involved with specific RNA–RNA interactions. U1 binds with the 5′ end of the intron and U2 near the 3′ end; U4 and U6 RNAs interact to form the U4/U6 complex, but once this RNP complex along with the U5 snRNP enters the super complex, the U6 binds to the U2 snRNA and the first stage of catalysis occurs. Figure 1 shows the general splicing cycle with special emphasis on the RNP complexes.

There is increasing evidence that the cleavage and ligation of the mRNA transcript are RNA-catalyzed reactions (Butcher and Brow 2005; Valadkhan 2005). Key components of this cycle occur in widely diverging eukaryotes, including the deeply branched protist, *Giardia lamblia* (Chen et al. 2007), confirming the earlier interpretation that a complete spliceosome was present in the last common ancestor of eukaryotes (Collins and Penny 2005). Even the minor spliceosome now appears to be an ancient eukaryotic feature (Russell et al. 2006; Davila Lopez et al. 2008).
There are tight linkages between splicing and the other mRNA-processing machineries (reviewed in Collins and Penny 2009). For example, in humans, transcription by RNA polymerase II and mRNA splicing are carried out in close physical proximity (Kornblith et al. 2004; Hicks et al. 2006), and spliceosome assembly also occurs before termination of transcription (Gornemann et al. 2005; Lacadie and Rosbash 2005). Furthermore, a number of splicing proteins, including U2AF (Millevoi et al. 2006), are known to interact with the complex that polyadenylates mRNA (Reed 2000; Millevoi et al. 2006). Splicing, 5′ capping, and 3′ polyadenylation are linked with other RNP-mediated functions. In a mammalian example, the exon junction complex (EJC) is deposited on the ligated mRNA during the second step of the splicing cycle and remains bound to the spliced mRNA as it is exported to the cytoplasm mediated by the Trex protein complex (Reed and Cheng 2005; Kerenyi et al. 2008). The EJC also contains several proteins involved in the nonsense-mediated decay (NMD) of incorrectly spliced or terminated mRNAs. Thus, the EJC set of proteins (bound during splicing to a transcribed mRNA) itself forms an RNP (mRNA + EJC), linking splicing to both RNA export and decay. A recent study (Jaillon et al. 2008) also shows that small introns from species as diverse as plants, animals, fungi, and the ciliate Paramecium tetraurelia are under strong selective pressure to cause premature termination of translation after incorrect intron retention (i.e., a splicing error), further linking, splicing, and NMD.

However, mRNA transcripts not only require modifications (i.e., splicing, capping, and tailing) but also require regulating. RNA interference (RNAi) (Munroe and Zhu 2006; Zhang et al. 2007) is well known for its role in regulating mRNA levels by either stimulating ( Vasudevan et al. 2007) or inhibiting ( Zhang et al. 2007) transcription or by mRNA removal. RNAi is also directly involved in many cellular processes including chromatin-mediated silencing and DNA rearrangements. The processing pathway for one RNAi mechanism microRNAs (miRNAs) shows RNP-like qualities. Here, double-stranded precursor pri-miRNAs are processed to pre-miRNAs in the nucleus by the Drosophila protein before they are exported to the cytoplasm where they are processed further by the Dicer protein (Zhang et al. 2007). Within this pathway, there are RNA–protein and RNA–export sectors, until the final regulatory action by the miRNA on the mRNA. There are many examples including small interfering RNA and PIWI-interacting RNA processing, but it is clear that as a general mechanism RNAi is an RNP-based form of regulation. RNAi, once thought a property of multicellular plants and animals, has been characterized recently not only in single-celled algae ( Molnar et al. 2007) but also in distantly related protists such as G. lamblia ( Chen et al. 2009; Prucca et al. 2008) and ciliates ( Lepere et al. 2008). Thus, RNAi as a general mechanism is likely to be another RNP-based process that occurred in the ancestral eukaryote.

Do all functional RNAs require protein cofactors to operate? Riboswitches are perhaps an example for which this might not be the case. Riboswitches are a class of RNA-based regulators where binding of a small metabolite, such as a cofactor, alters the tertiary structure of mRNA and downregulates its own expression ( Montagne and Baty 2008). Riboswitches are well known in prokaryotes, but their role in eukaryotic gene regulation is only now becoming clearer. Thiamine pyrophosphate (TPP) riboswitches (responding to TPP, a derivative of vitamin B1) have now been reported in both fungi ( Kubodera et al. 2003; Cheah et al. 2007) and green algae ( Croft et al. 2007). Once thiamine binds to an intron, it leads to alternative splicing resulting in an inactive mRNA and thus preventing the biosynthesis of additional thiamine. It is unclear whether thiamine biosynthesis is ancestral in eukaryotes or, for example, in the plant lineage it was introduced with the chloroplast as a step toward autotrophy. This would imply that the alternative splicing for thiamine biosynthesis in plants was a secondary addition to this riboswitch. However, alternative splicing itself does appear to be quite old in eukaryotes in that some sites are conserved between plants and animals ( Irimia et al. 2008).

The final point here, and as predicted from problems associated with molecular crowding, is that the large number of protein interactions within RNP-based processes applies to other proteins that do not interact directly with RNA. There are well-established databases of protein interaction networks (PINs) such as BioGrid ( Stark et al. 2006; Breitkreutz et al. 2008) or iHop ( Hofmann and Valencia 2004) that can be used to quantify the extent of interactions between proteins. But even these large networks only represent a fraction of the true interactions taking place within an organism as often multiple isoforms created by alternative splicing are ignored ( Stumpf et al. 2007). So let us say, as an example, we choose from the BioGrid v2.0.35 data set 6 key proteins in the EJC and the Trex complexes and examine their interaction with other proteins. We find only 2 subnetworks result indicating a high connectivity between the EJC and Trex proteins ( Figure 2A) and ask if this level of connectivity is larger or smaller than expected for interactions between groups of 6 proteins selected randomly from the same original data set? Figure 2B shows that by chance we do expect to see more than 2 subnetworks. That is, from 6 randomly selected proteins we expect to see more unconnected subnetworks than is found for the 6 EJC and Trex proteins. Thus, we observe a more integrated network of protein–protein interactions than expected just from random interactions between proteins.

This is only a small example looking at a small section of the interactome, and so global network properties like overall connectivity cannot be applied here. However, larger studies summarized in Stumpf et al. (2007) have been somewhat inconclusive due to the requirement for improved models of network evolution. Using the small example above, our point is that the physical location and functional interactions of these particular proteins favor network formation ( Ivinic et al. 2008a; Ivinic et al. 2008b; Kelly and Stumpf 2008; Levy and Pereira-Leal 2008).
However, we should point out that PINs are averaged structures collating interactions across the entire life of a cell (in contrast to the dynamic interactions of the RNA infrastructure). Combining these 2 types of networks for an understanding of the RNP infrastructure will be difficult as it necessitates the moving from averaged PINs to dynamic PINs where protein interactions also change.

At this point, our principle conclusion is that the RNP infrastructure is extensive and integral to the mediation of gene expression; it must occupy a major place in our search for the origins of the eukaryote cell.

**RNPs as Nuclear Compartments**

The highly complex structure of the eukaryote cell is sometimes considered a “problem” that can be “explained” by the origin of eukaryotes by one of the fusion models. However, this need not be so. The efficiencies with which RNP complexes mediate the individual steps of the gene expression cascade depend on having all the components together when and where they are needed. Traditionally, the cell was regarded as simply “a bag of enzymes,” but when we understand that the availability of proteins depend on the diffusion of macromolecules to a working site we see that the ensuing delays would be costly (e.g., Ehrenberg and Kurland 1984). As mentioned above, such delays would be accentuated by the densely packed cytosol of cells that reduces free diffusion of macromolecules (“molecular crowding,” Ellis 2001; Kurland et al. 2006). In contrast, association of the components into complexes secures their availability for the steps within that cycle.

Each round of gene processing (e.g., the splicing of multiple introns) requires the cooperation of hundreds of components to mediate their functions. Accordingly, this processing can only function efficiently by integrating the components into extensive macromolecular complexes. In effect, spliceosomes (along with nucleoli, Cajal bodies, gems, and speckles) are important examples of the ubiquitous tendency toward complex formation. These nuclear compartments are in fact RNP associations that minimize the delay times for individual steps of gene processing.

The retardation of diffusing macromolecules is exacer-
bated in eukaryotes by their large cell volumes, much larger on average than archaea and bacteria. The localization of complexes within the nuclear, nucleolar, endoplasmic, and cytoplasmic compartments enhances the effectiveness of gene processing by maximizing efficient dynamics (Kurland et al. 2006). In general, macromolecular complex formation and subcellular partitioning are structural adaptations that solve kinetic problems (Kurland et al. 2006). Thus, we see, based on physicochemical principles, that the highly structured and compartmentalized nature of the eukaryote cell is sometimes considered a “problem” that can be “explained” by the origin of eukaryotes by one of the fusion models. However, this need not be so. The efficiencies with which RNP complexes mediate the individual steps of the gene expression cascade depend on having all the components together when and where they are needed. Traditionally, the cell was regarded as simply “a bag of enzymes,” but when we understand that the availability of proteins depend on the diffusion of macromolecules to a working site we see that the ensuing delays would be costly (e.g., Ehrenberg and Kurland 1984). As mentioned above, such delays would be accentuated by the densely packed cytosol of cells that reduces free diffusion of macromolecules (“molecular crowding,” Ellis 2001; Kurland et al. 2006). In contrast, association of the components into complexes secures their availability for the steps within that cycle.

Each round of gene processing (e.g., the splicing of multiple introns) requires the cooperation of hundreds of components to mediate their functions. Accordingly, this processing can only function efficiently by integrating the components into extensive macromolecular complexes. In effect, spliceosomes (along with nucleoli, Cajal bodies, gems, and speckles) are important examples of the ubiquitous tendency toward complex formation. These nuclear compartments are in fact RNP associations that minimize the delay times for individual steps of gene processing.

The retardation of diffusing macromolecules is exacer-
bated in eukaryotes by their large cell volumes, much larger on average than archaea and bacteria. The localization of complexes within the nuclear, nucleolar, endoplasmic, and cytoplasmic compartments enhances the effectiveness of gene processing by maximizing efficient dynamics (Kurland et al. 2006). In general, macromolecular complex formation and subcellular partitioning are structural adaptations that solve kinetic problems (Kurland et al. 2006). Thus, we see, based on physicochemical principles, that the highly structured and compartmentalized nature of the eukaryote cell.
cell is simply a necessity. This does not mean that the
eukaryote cell did not arise by some fusion process creating
some of these compartments and only that the compart-
mentalized nature of the eukaryotic cell is, in itself, not an
argument for or against such a process.

In addition, RNP complexes may be essential to the
stability of both RNA and protein components in a cytosol
policed by nucleases and proteases. It has been known for
some time that posttranslational destruction of aberrant
polypeptides eliminates proteins that lack the normally
compact domain structures (Goldberg and Dice 1974;
Glickman and Ciechanover 2002). This depends on
recognition of unfolded sequences in abnormal proteins.

Opposing the proteolytic complexes are batteries of
chaperonins that also recognize abnormal sequence folds
in proteins to mediate a refolding process that may repair
an abnormal structure (Wickner et al. 1999; Gidalevitz et al.
2006). The tug of war between chaperonins and proteolytic
functions determines how effectively a population of
abnormal proteins is eliminated from the cytosol (Wickner
et al. 1999; Maisnier-Patin et al. 2005). A bound ligand, such
as an oligonucleotide sequence (i.e., RNA), may tip the
balance in favor of stability. Once bound to a protein
domain, RNA will protect that domain by excluding it from
most proteolytic sites because these only accommodate
amino acid sequences in a slim beta fold (Tyndall et al.
2005).

Similarly, all cells are invested with a gauntlet of
nucleolytic enzymes that make short work of unprotected
RNA molecules. In eukaryotes, these nucleolytic com-
plexes may be as large and as elaborate as exosome
vesicles (reviewed in Vanacova and Stefl 2007). This may
be an important reason why we do not observe RNAs
unaccompanied by protective protein partners. In addition,
proteins also function with RNA the way that
chaperonins do with proteins, that is, by favoring particular functional foldings in RNA. Furthermore,
a bound oligonucleotide gives a protein an ideal ligand
with which to recognize a complementary nucleotide
sequence. Such nucleotide sequences recruited as protein
cofactors are ideal for sequence-specific recognition of
binding sites in RNA.

In contrast to eukaryotes, we do not find a comparably
rich flora of small RNP complexes in the cascade of archaea
and bacteria. What we find in these microbes is a bare bones
cascade in which DNA codes RNA that makes protein.
Missing in prokaryotes is the complexity for the more
elaborate infrastructure that provides the regulation, trans-
port, storage, and garbage-handling functions associated
with eukaryote cell compartmentalization. However, we do
have evidence of some RNP infrastructure within prokar-
yotes with the identification of an apparently protein-free
RNA domain that mediates peptide bond formation in
bacterial ribosomes (Nissen et al. 2000). It has been found
that protein L27 as well as other proteins are essential for
peptide bond formation by bacterial ribosomes (Maguire
et al. 2005), with the apparent absence of protein from the
putative peptidyl transferase domain attributed to the

The RNP Dreamtime—Life before DNA

So far, this overview has focused on the network of RNP
complexes in modern cells, but clearly the question of the ultimate
origin of the RNP infrastructure is fundamental, as is the
relationship between eukaryotes and the 2 groups of
prokaryotes (archaea and bacteria). Given the highly
superior properties of proteins as catalysts (relative to
RNA, Jeffares et al. 1998), the simplest hypothesis is that the
basic RNP infrastructure has been inherited from an RNA–
protein world; a world before DNA was the main coding
molecule (Penny and Poole 1999; Poole et al. 2002; Forterre
2005). Because this is so far back in time, we paraphrase it as
the “RNA dreamtime”—a time for which we have only
rough glimpses from modern cells.

Given the standard view that RNA genomes preceded
DNA genomes (Joyce 2002), at least one potential route for
the evolution of modern splicing is apparent. The rates of
evolution for genomes encoding short polypeptide sequen-
ces, as well as their longer protein derivatives, will improve
with recombination between different partial sequences in
a manner similar to that which occurs in RNA viruses in the
present. For this reason, it is likely that recombination
would emerge in a world of RNA genomes (see Lehman
2003; Egel and Penny 2007). During the transition to DNA
genomes, such functions could have been recruited as
splicing machinery for RNAs as one way to support a major
expansion of novel proteins compounded from a basic set
of protein fold superfolds in eukaryotes (see Kurland et al.
2007). This is certainly consistent with our suggestion that
some modern RNP complexes are relics from the RNA
dreamtime.

Are the major RNP classes involved in the RNA
infrastructure ancient or did they arise only in some lineages
of modern eukaryotes? Four hypotheses seem relevant to
this issue.

1. There is continuity between both the major classes of
modern RNAs, and their proteins, back to the RNP
world. This is accepted for tRNA, mRNA, and rRNA
and for rRNA modifications (uracil to pseudouracil and
the 2-O-methylation of ribose). Is it more general,
including RNAi?
2. New RNAs arise but function with preexisting proteins. We expect that such cases occur simply because so much RNA is transcribed in the cell that some small RNAs must occasionally be recruited, augmenting existing functions.

3. Genuine novelties may arise where both the RNA and the protein components arise de novo (either from a new protein see Keese and Gibbs [1992] or recruitment from a duplicated protein that originally had a different function). We do not know of any such cases yet, but we certainly should be alert to them, and we would be surprised if some did not occur.

4. The complexities of the RNP system occurred in the RNP world and were then lost in early prokaryote evolution and then later by "magick" reappeared in eukaryotes. Given the major advantages of proteins over RNA for catalysis (Jeffares et al. 1998), it seems unlikely that RNA would ever "take back" a function that proteins were already carrying out.

Aspects of these hypotheses are (in principle) testable, but we will need more knowledge of the distribution of each type of RNP complex in eukaryotes. Considering the last hypothesis, if archaea and bacteria did not presently exist, then the first hypothesis (RNA continuity) would be the obvious null hypothesis with continuity from the RNP world to the modern DNA genome world (Figure 3).

There are other important issues that may be related to earlier phases of the RNP-to-DNA genome transition. Though prominent in gene expression, RNP complexes are scarce in membranes and essentially absent from the biochemistry of small molecules (although many enzyme cofactors are related to ribonucleotides [White 1976]). Not finding RNA in small molecule biochemistry is informative (see Discussion in Jeffares et al. 1998). Were peptides or proteins preferentially exploited in these domains (White 1976) even in the RNA dreamtime? Indeed, recent experimental work shows affinity between some RNA molecules and lipid bilayers (Janas et al. 2006), and many experiments have been undertaken on catalysis by RNA and cofactors (Jadhav and Yarus 2002; Joyce 2002). It is difficult to consider meaningful scenarios for the evolution of gene processing without assuming that peptides enhanced the fitness of even the most ancient RNA-based cellular systems. Eventually, we will want to follow the continuity (Penny 2005) as illustrated in Figure 3, from a strictly biochemical phase to an ancient RNA world to the RNP world and then to DNA being used for the storage of information. At the moment, it is exciting to be able to infer possible processes from just before the transition to DNA genomes.

**Conclusions**

Any understanding of the origin of the eukaryote cell must consider the recent harvest of information about the involvement of RNP complexes in RNA expression, regulation, and modification in eukaryotes. If we take the traditional view of an early pre-DNA world leading to prokaryotes, and then to eukaryotes, then virtually nothing is learned about the critical stage of the RNP world during the origin of life. It is just not feasible to base eukaryotic origins on cellular and gene processing properties of modern bacteria and archaea, and it is time we emphasized the use of eukaryotes as models of eukaryotic evolution. Furthermore, if we explore the RNA continuity hypothesis (Figure 3A), we can investigate the RNP infrastructure evolved in the RNP to DNA transition. Eukaryotes, rather than presenting the pinnacle of advanced cellular evolution, indeed provide excellent opportunities to understand more about the later stages of the origin of life, the transition from RNA genomes to DNA genomes.

**Funding**

Allan Wilson Centre for Molecular Ecology and Evolution, Institute for Molecular Biosciences, Massey University (L.J.C., P.B., D.P.); Royal Physiographic Society (to C.G.K.); Royal Academy of Sciences (to C.G.K.).
Acknowledgments

We are grateful to Geoff Jameson for help with the literature and for reliably informative conversation. We also thank our colleagues at the Allan Wilson Centre.

References


